DOCUMENT RESUME

ED 449 182 TM 032 290

AUTHOR Kim, Jong-Pil
TITLE Meta-Analysis of Equivalence of Computerized and P&P Tests on Ability Measures.
PUB DATE 1999-10-00
NOTE 45p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 13-16, 1999).
PUB TYPE Books (010) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Effect Size; *Meta Analysis; *Scores; *Statistical Significance; Synthesis
IDENTIFIERS *Paper and Pencil Tests; Type I Errors

ABSTRACT

This study was conducted to investigate the equivalence of scores from paper-and-pencil (P&P) tests and computerized tests (CTs) through meta-analysis of primary studies using both kinds of tests. For this synthesis, 51 primary studies were selected, resulting in 226 effect sizes. The first synthesis was a typical meta-analysis that treated multiple measures from the same subjects within studies as independent data. The second synthesis represented results using composite effect sizes. The results from both syntheses were compared in terms of grand mean effect size and the findings for moderator variables. The results of one analysis indicate that eliminating dependence between equivalent scores does not affect the significance of homogeneity tests very much. Overall, ignoring non-independence between equivalent scores tends to lead to underestimated standard errors and inflated Type I error rate when determining statistical significance tests. This is not always true, however, because the means, dispersions, and distributions of equivalent scores depend partly on the number of equivalent scores and partly on the methods for adjusting for dependence of equivalent scores. The type of computerized test was the most important variable when evaluating the equivalence between CT and P&P tests. For computer adapted tests, mathematics, source, and possibly sampling age are significant variables, but for computer based tests, the analyses did not find a significant moderator. (Contains 11 tables, 1 figure, and 78 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Meta-analysis of Equivalence of Computerized and P&P Tests on Ability Measures

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Kim, J.-P.

ED 449 182

JONG-PIL KIM

Michigan State University

TM032290

* Paper presented at the annual meeting of the Mid-Western Educational Research Association, 1999, Chicago, IL.

BEST COPY AVAILABLE

With the rapid development of the computer and of item response theory (IRT), computerized tests (CTs) have been widely applied. Even though a variety of CTs have been used, doubt continues about the equivalence of the scores from paper-and-pencil (P&P) tests and CTs, largely based on the different modes of the two tests. In other differences in the testing environments and the administrative modes for P&P and computerized tests may affect the individual examinees in some way. This study was conducted to investigate the equivalence of scores from P&P tests and CTs via meta-analysis on primary studies, which had used both P&P tests and computerized versions of P&P tests. In addition, the effect of nonindependence of effect sizes on the equivalence of the test forms was investigated.

Theoretical Framework

Computerized Testing

Computerized tests (CT) may be divided into two major categories, computerized adaptive tests (CAT) and computer based tests (CBT). A CAT is one in which different sets of test questions (items) are administered to different individuals depending on each individual's status on the trait being measured (Weiss, 1985). Considering the responses of the examinee on the previous item(s), additional items are selected from an item pool with items of known difficulty and discrimination. Thus, not all examinees receive the same set of test items. In contrast, CBT generally refers to the use of computers to administer a conventional (that is, P&P) test. As a result, all examinees receive the same set of test items.

Understanding CBT is easy because the components are just the same as those in traditional tests, except for using the computer mode. However, a CAT has much different components than either a P&P test or CBT. Weiss and Kingsbury (1984) summarize the main components of a CAT as (a) an item response model: one-, two-, or three-parameter IRT model, depending on the nature of the items used and the fit of the item responds data to the model chosen; (b) an item pool with estimated item parameters: difficulty levels of items in the pool must span the full range of trait levels in the population; (c) an entry level, chosen according to each student's ability level; (d) an item selection rule: maximum information or Bayesian; (e) a scoring method - maximum likelihood or Bayesian; and (f) a termination criterion: a rule for ending the test, prior to test administration.

CA testing strategies have been designed to utilize item information data (e.g., Brown & Weiss, 1977; Maurelli & Weiss, 1981; Weiss & Kingsbury, 1984). For instance, the maximum information adaptive testing strategy selects items that provide maximum levels of item information

at an individual's currently estimated trait level. In addition, IRT-based methods of scoring tests permit estimation of individuals' trait levels based on their responses to one or more items. As a consequence, an item can be administered and an estimate can be made of the individual's level on the trait. After the administration of an item and estimation of the trait level, the new trait level is used to select the next item to be administered to that examinee to provide maximum information for the current estimated level of the trait (Weiss, 1985).

With the Bayesian method, each examinee begins the test with an initial trait-level estimate and a confidence interval associated with that estimate. These are operationalized as a mean and variance of a normal prior distribution on the trait being measured. As each item is answered, a new trait estimate is calculated using the response and the prior distribution values, and a posterior distribution of trait estimates is developed. The Bayesian selection method chooses the item that most reduces the Bayesian posterior variance. Specifically, the posterior variance is calculated for every available item in the pool, given the candidate's current trait estimate and the item's parameters. The question that reduces the posterior variance to the smallest value is chosen (Vispoel & Coffman, 1994; Olsen, Maynes, Slawson, & Ho, 1986).

The mathematical model that guides the adaptive testing process provides a scale, referred to as the proficiency or θ scale. Any test that is composed of items that have been fit by some IRT model produces scores on the proficiency scale. This is true for conventional P&P tests as well as CATs. The difference between the two types of tests is that adaptive tests require the proficiency scale or some derivative thereof during item administration, whereas conventional tests can manage with a simpler scale, such as number right. Adaptive tests require a scale that is not tied into a particular set of items because adaptive test scores are based on many different item sets.

Test Equivalence

It is generally agreed that before an assessment developed from an existing P&P version is adapted for computer administration, the equivalence of the two forms needs to be adequately demonstrated. To establish equivalence, it must be demonstrated that both versions of the test yield the same score, or at least parallel scores. Guideline 16 of the American Psychological Association's Guidelines (The American Psychological Association, 1987) for CTs states that (1) the equivalence scores from CT versions should be established and documented before using norms or cutting scores obtained from conventional tests to interpret scores from the CT versions of conventional tests, and (2) the equivalence may be held if (a) the rank orders of scores of individuals tested in alternative

modes closely approximate each other, and (b) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode.

Some purists question whether CATs can ever be equivalent with conventional tests because each examinee's test has a different number of items that may also differ in level of difficulty. But if both tests are measuring the same construct, which has been thoroughly demonstrated in the case of CAT-ASVAB (Greaud & Green, 1987; Green, 1987; Moreno, Wetzel, McBride, & Weiss, 1984; Vicino & Hardwicke, 1984), then the two scales can be compatible. If the same proficiency is being assessed, if samples are selected to be representative of the intended test-taking population, if common equating items are in fact measuring the same thing, and if an appropriate equating model is employed, then it should be possible to correctly equate the scores produced by an adaptive item pool to other tests or item pools.

The most serious of the potential unintended consequences of CT is the possibility that it may disadvantage some groups of test takers (Power & O'Neil, 1992). The Office of Technology Assessment of the U.S. Congress (1992) also pointed out that inequity may arise in the context of computer-based assessment to the extent that test taking involves procedures with which not all test takers are equally comfortable. These concerns with equity issues started with the fact that not all persons have similar experience in using computers (Green, Bock, Humphreys, Linn, & Reckase, 1984). Haney (1991) stated the importance of not harming people in testing. Current emphasis on testing and the importance attached to test results places a special responsibility on educators to use testing methods that provide valid and reliable information without harming students or disrupting the educational program. As Haney implied, even if CT has a lot of advantages including higher reliability, efficiency, and convenience, it should not be accepted as a good testing method in educational situations with equity problems. It is necessary to determine whether or not certain groups of people may be adversely affected by a CT process (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978).

Based on a wide literature (Lee, 1986; Llabre, Clement, Fitzhugh, & Lancelota, 1987; Heinssen, Glass, & Knight, 1987; Martinez & Mead, 1988; Wilder, Mackie, & Cooper, 1985; Lockheed, 1985; Ward, Hooper, & Hannafin, 1989; Fletcher & Collins, 1986; Wise & Plake, 1989; Lunz, Bergstrom, and Wright, 1992; Vispoel, Wang, de la Torre, Bleiler, & Dings, 1992; Stone & Lunz, 1994; Mazzeo and Harvey, 1988; Wise and Plake, 1989, 1990; Kovac, 1990; Wainer & Kiely, 1987), the question of equivalence is often raised because the mode of administration of the

CTs differs from that of the P&P test: are the scores obtained via two different modes of administration the same even if appropriate equating procedures are implemented? The first purpose of this study is seeing if the scores in P&P tests are equivalent with the scores in CTs.

Review of two previous meta-analyses

Two previous meta-analyses of CT (or CAT) and P&P tests of ability measures (Bergstrom, 1992; Mead & Drasgow, 1993) were done 6 or 7 years ago. Since 1993 even more investigations of the equivalence of CT and P&P tests have been done (at least 10 studies with 58 effect sizes were found). Additionally, the previous meta-analyses did not include dissertations, which often report well-designed research. A more up-to-date meta-analysis is thus needed to accumulate new studies in this area. Also, even though many studies have applied CT to classroom examinations (11 studies with 45 effect sizes were found), few of these studies were synthesized. The previous meta-analyses focused on tests of achievement and cognitive ability, respectively. However, the terms, aptitude, ability and achievement may be equivalent functionally. Bond (1989) wrote, "Cooley and Lohnes (1976) have in fact claimed that the distinction is a purely functional one. If a test is used as an indication of past instruction and experience, it is an achievement test. If it is used as a measure of current competence, it is an ability test. If it is used to predict of forecast future performance, it is an aptitude test" (p. 429). For this meta-analysis, research using any of the three tests is included.

Bergstrom (1992) reported a grand mean effect size of – .002 between CAT and P&P tests in achievement measures 15 effect sizes, which was not significant. She examined one moderator variable, the effect of administration order. When significant differences were found mean measures were higher for a pre-existing P&P test than a post-existing CAT when the same examinee took both. Mead and Drasgow (1993) reported a .91 correlation across administration modes of CT and P&P tests. They found no significant difference between CT and P&P for power tests (a mean of r = .97 from 123 correlations), but found one for speed tests (with mean of r = .72 from 36 correlations). This implies that modes of administration affect the equivalence of speed tests, but when examinees are given sufficient time to solve items, there is no mode effect. Moreover, CTs were found to be slightly more difficult than conventional tests. Mead and Drasgow attribute the effect on speeded tests to differential motor skills that are required in conventional as compared with computerized testing. In addition, they report that four moderators were significant, namely, use of random assignment, differential motivation (why the examinees took the tests), sample size, and type of report (journal and presentation vs. technical report and manuscript) in predicting the equivalence

scores of CTs and P&P tests. On the other hand, they reported that the method of administration of the computerized version and publication year were not significant moderators.

One particular focus in the Mead and Drasgow study is their consideration of speededness in test. In a pure power test, the items range in difficulty and there is no time limit. The goal is to measure how accurately the examinees can answer the items. In a pure speed test, the items are very easy and the time limit is very strict. The goal is to measure how quickly the examinees can answer items. In reality, most tests contain both speed and power components, and these are called speeded tests. Speeded tests usually result from administering a power test with a time limit, a practice that is often required when the test is group-administered (Schnipke, 1995). More importantly, speededness is a problem for IRT. Unidimensional IRT implicitly assumes that the test is unspeeded; speed would be another dimension (Hambleton & Swaminathan, 1985). When estimating IRT item parameters on a simulated speeded test, the *a* and *b* parameters tend to be overestimated and the *c* parameters underestimated for the items toward the end of the test (Oshima, 1994). Thus CAT studies, which tried to develop tests to be equivalent to P&P speed tests are not synthesized in this study.

The previous meta-analyses focused on article characteristics and study characteristics as moderators of CAT/P&P differences. However, some studies have examined individual differences in CT situations including gender, anxiety, computer experience, ethnicity, and motivation. Examinee sample characteristics might be interesting moderator variables in this synthesis because the mode of test administration may interact with individual differences characteristics. Additionally, test characteristics such as subject area (test content) and test type (standardized battery vs. classroom examination) may affect the equivalence. Finding variables that moderate the difference between CT and P&P tests is the second purpose of this study.

Nonindependence Issue

Landman and Dawes (1982) cautioned about five sources of nonindependence in meta-analysis. First, they cite multiple measures of outcomes from the same subjects within single studies; second, measures taken at multiple points in time from the same subjects (i.e., multiple occasions); third, nonindependence of scores within a single outcome measure; fourth, nonindependence of studies within a single article; and fifth, nonindependent samples across articles (p. 506-507). The third source appears when a study reports both a global index as well as more specific index, which is a part of the global index. In this case, choosing the specific index is ideal if it allows the study of interesting moderator variables. The fourth type of dependence occurs when samples from two

different experiments reported in a study are overlapping or the same. The last type of dependence appears if the same sample appears in two different articles. In this synthesis the more informative article was selected.

The first type of dependence is common in studies of CT and P&P tests. Nineteen of the 50 studies in the current synthesis report more than one outcome measure. The typical ad hoc analysis may treat each effect size from a given study as independent of the other effect sizes from the same study (e.g., Smith, Glass, & Miller, 1980). However, Glass, McGaw, and Smith (1981) recognized that "the data set to be analyzed [in a meta-analysis] will invariably contain complicated patterns of statistical dependence [since] each study is likely to yield more than one finding" (p.200). Bangert-Drowns (1986) stated, "multiple effect sizes from any one study cannot be regarded as independent and should not be used with statistical tests that assume their independence" (p. 397). In the same article (p. 392), he discussed the "Inflated Ns" problem. A report will have a greater influence on the meta-analytic findings if it continues many dependent measures. The "Inflated Ns" problem threatens the generalizability or external validity of a meta-analysis. Another problem is inflated Type I error (Raudenbush et al., 1988). Strube (1983) mentioned a general rule, that is, failure to adjust for nonindependence inflates the Type I error rate at the meta-analysis level.

Researchers have devised several methods for combining dependent data in meta-analysis. A strategy for reducing dependence of data is to select, on some predetermined basis, a single dependent measure to represent each study (Cooper, 1979). But, the question "what is the best indicator among several dependent variables?" is too ambiguous. It is very difficult to make such a decision. A common strategy for dealing with studies that use multiple outcomes has been to average. This makes sense for providing a representative effect size estimate when the outcomes are parallel measures of a single construct (Raudenbush et. al., 1988). Instead of the mean, the median effect size is a more conservative option.

[A similar, more sophisticated solution proposed by both R&R (1986) and Olin & Glaser (1994) is to create a weighted composite of the multiple effects for each study. In this research I examine the use of O&G's composite to deal with dependence in the CT/P&P studies.]

A statistical solution for this nonindependence problem within a study has been developed by Rosenthal and Rubin (1986). When the study has a big sample size and small differences of the intercorrelations between outcome measures, they suggest computing a composite effect size. Gleser and Olkin (1994) also showed how to calculate composite effect sizes within studies by using all individual intercorrelations among outcome variables. One difference between these two procedures

is that Rosenthal and Rubin (1986) use a "typical" correlation, which is a correlation representative of all intercorrelations between the multiple measures. Thus this investigation focuses on Glaser and Olkin (1994) calculation because of its relative accuracy. In this synthesis a grand mean effect size from a typical meta-analysis (one that treats effect sizes within studies as independent) is compared with the grand mean effect size based on composite effect sizes by Gleser and Olkin procedures).

In summary, the three purposes of this study are to (1) update earlier meta-analyses with most recent findings on the equivalence of CT and P&P tests for ability measures, (2) examine the influence of moderators (characteristics of studies, samples and tests) on test equivalence, and (3) investigate the impact of within-study dependence on the overall effect size(s) and analyses from the synthesis.

Methods

Literature Retrieval

Primary studies were selected using four criteria: a) the study provided sufficient information for computing an effect size (i.e., means and standard deviations of two groups for CT and P&P tests or other information like *r*s (correlations), *t*-statistics, or *F*-statistics which can be transformed to an effect size *d*), b) the tests measured abilities, achievement, or aptitude, c) the within-group sample sizes were greater than 10 and were not seriously unbalanced (no less than 40% could be in one subgroups), and d) if the same samples were analyzed in different articles, the more informative study was selected to avoid nonindependence across articles (Landman & Dawes, 1982).

Finding the studies from the two previous meta-analyses was the first step in my literature search. All eight studies from Bergstrom (1992) were available including three of Bergstrom's own copies. Fifteen studies from Mead and Drasgow's (1993) research synthesis were found. However, the other 14 unpublished studies could not be obtained. Three more studies were identified in Neal (1991) which presented a brief summary of 11 references concerning CT compared with P&P tests.

The whole process of selecting studies from the Dissertation Abstracts Data Base was done in one sitting by using as keywords "paper-and-pencil test" or "conventional test" along with either "computerized test," "computerized adaptive test," "computer based test," and "computer assisted test" with "ability" or "achievement." Ten dissertations were identified. Since all dissertations reported the standard deviations and means for CTs and P&P tests in some way, all dissertations are analyzed. The ERIC (Educational Resources Information Center) electronic data base and PSYC

data base for psychological journals in the Michigan State University library were also searched in the same manner described as above, and 34 additional studies were identified. Thirteen of these studies were removed from this analysis because they did not include sufficient information to compute effect sizes. In addition, 3 studies were eliminated because they had a sample size of less than 10 or were seriously unbalanced. If the same study appeared as both a journal article or a dissertation and as an ERIC document, the dissertation or journal article was selected (3 studies were removed here). If the same sample appeared in two different studies, the study with interesting moderators was selected to avoid nonindependence across articles (1 study was removed for this reason).

As a result, 51 primary studies were selected for this synthesis. The primary studies are listed in Appendix A. A descriptive summary of the 51 primary studies is presented in Table 1. Most of these studies have been conducted since 1989 or with college student or adult examinees. The fact that so many of this research involves either studies on classroom tests (30.7%) or dissertations (21.2%) is significant for this synthesis because the previous meta-analyses did not include either source.

Table 2 summarizes the characteristics of 226 effect sizes from 51 studies. The percentage of effect sizes from computer based testing studies is around 66%. English and Mathematics tests are used in more than half (52.6%) of the studies. This indicates that efforts for computerized testing have been primarily devoted to these subjects. The studies using young students (below high school age) are just 4%, which suggests that there are some restrictions to using computers to test younger experiences. There are only 15 effect sizes (6.6%) were based on nonrandom samples. Under design characteristics, random refers to studies using random equivalent group design; "P&P 1st" means that the examinees took a P&P test before taking the CT version, and similarly "CT 1st" means the examinees took a CT before taking the P&P test.

Coding Sheet and Coding Procedure

Data related to four overall areas were coded, namely, article characteristics (type of publication, name of source, and publication year, etc.), sample characteristics(grade level, number of examinees who took a particular test and total sample size, etc.), study characteristics(which characteristics consist of design aspects which ask whether the sampling is random or nonrandom, and whether all samples took both modes of the test), and test characteristics(test name, type of computerized test, and subject area of the test, etc.). The author coded all of the primary studies.

Eight doctoral students who had experience in implementing meta-analysis or had taken a meta-analysis class volunteered to code 6-7 primary studies each. The percentage of agreement between the author and the other coders is calculated by treating all other coders as if they are a single coder. Agreement percentage between the author and the other coders was 100% for type of source, source name, publication year, and sample, 90% for total sample size, 88% for category of computerized test (CAT or CBT), 80% for study design and review of CT, 84% for name of test, and 65% for speededness. The average agreement was 88.7%.

Analyses

Two steps were implemented in analyzing the primary studies for this synthesis. Synthesis I represents a typical meta-analysis which treats multiple measures from the same subjects within studies as independent data. Synthesis II represents results using composite effect sizes. The results from synthesis I and II are compared in terms of grand mean effect size and the findings for moderator variables.

Synthesis I

The effect size computed is the standardized mean difference between the achievement measure estimated by the CT and the achievement measure estimated by the P&P test. The formula $(\bar{X}_{iCT} - \bar{X}_{iP\&P})/S_i$ is used to calculate the biased effect size (d_i) for each study, where $\bar{X}_{iCT}$ is the mean achievement measure on the CT, $\bar{X}_{iP\&P}$ is the mean achievement measure on the P&P test and S_i is the pooled standard deviation for study i calculated as:

$$S_i = \sqrt{\frac{(n_{iCT}-1)(S_{iCT})^2 - (n_{iP\&P}-1)(S_{iP\&P})^2}{n_{iCT}+n_{iP\&P}-2}}, \qquad (1)$$

where n_{iCT} is the number of examinees who took the CT and $n_{iP\&P}$ is the number of examinees who took the P&P test (Bergstrom, 1992, p.8). The unbiased effect size, conditional variance, and homogeneity test are implemented based on Hedges & Olkin (1985). To find if there is difference between subgroups and if each subgroup is heterogeneous, omnibus tests for between-groups differences and for within-group variation in effect are implemented.

General least square regression is implemented to see which moderator variables of interest predict the effect size or equivalent scores (ESs). All tests for the regression were implemented based on Hedges & Olkin (1985).

Synthesis II

In synthesis II, the composite effect sizes are calculated by Gleser & Olkin method. Gleser & Olkin (1994) showed how to obtain composite effect sizes when outcome variables are correlated. The composite effect size within a study is calculated using:

$$\hat{\delta}_i = \sum_{j=1}^{p} a_{ij} d_{ij}, \qquad (2)$$

where p is the number of effect sizes (or number of outcome measures) of study i, d_{ij} is the jth effect size in the ith study, and

$$(a_{i1}, \ldots, a_{ip}) = \left[\frac{1}{e' \psi_i^{-1} e} \right] e' \psi_i^{-1}, \qquad (3)$$

where e equals to (1, 1 , ..., 1)' and ψ_i is the variance-covariance matrix in study i. The variance of the composite effect size is given by $\left(e' \psi_i^{-1} e\right)^{-1}$ (Gleser & Olkin, 1994, pp. 352-353).

Not all studies report the intercorrelations between outcome variables. In such cases, missing intercorrelations were imputed from similar studies which report intercorrelations between the same outcome measures for similar samples. When study i has more than one outcome measure, the composite effect size $\hat{\delta}_i$ replaces the typical effect size d_i to compute the unbiased effect size and its conditional variance.

Results

Synthesis I

The Q statistic of the homogeneity test for all 226 effect sizes is 1226 ($p \leq .0001$, df = 225), which indicates heterogeneity of the effect sizes. When separated, 77 CAT ESs and 148 CBT ESs

are also heterogeneous. This finding supported use of a random effect model[1] rather than a fixed effect model for further analyses. The mean ES across all studies is .019, and a 95% confidence interval (CI) is −.03 to .068, indicating that even if the CT score on average is slightly higher than P&P (ES = CT − P&P), it is not statistically significant. However, the results are not all homogeneous, so this simple result does not tell the whole story. Table 3 summarizes the categorical analyses. A significant Q statistic between adaptive types indicates that there is a significant difference between the types of computerization, CAT and CBT. While CAT has a negative ES, CBT has a positive ES. For CAT, while the Q-between statistics for sample, sample size, and test type are not significant, the Q-between statistics for publication year, source, test type, content and design ($p < .05$) indicate significant differences between subgroups. From the individual 95% CIs, one can make the following conclusions: first, performance levels on CAT versions of standardized tests and classroom tests are not equivalent with those for P&P tests; second, CAT versions of mathematics and other cognitive tests (e.g., recognition, logical reasoning, etc.) appear equivalent with P&P tests.

For CBT, while the Q-between statistics for sample size and content are not significant ($p > .05$), the Q-between statistics for publication year, source, sample, test type, and design ($p < .05$) indicate significant differences. These results for CBT are the same as for CAT, except for the variables "sample" and "content." The ESs are equivalent for school-based examinees of college age and older, and those below high school age.

Regression analyses with a mixed effects model were implemented to evaluate moderators. The correlation between the predictor year and content is higher than .8. To avoid multicollinearity, the variable publication year was not included in the regression analyses because it is relatively less significant in measurement settings. For the mixed effects model, the variance of each data point is defined as v_i (from the fixed effect model) plus $\sigma^2_{\theta x}$. The estimate of $\sigma^2_{\theta x}$ is calculated from an approximation that mean square residual from the general regression model minus the estimated variance (mean of variances) (Raudenbush, 1994, pp. 310–311). For the model significance tests

---

[1] For the random effect model, the variance is defined as $v_i + \sigma^2_\theta$ where is v_i the variance from the fixed effects model. The estimate of $\sigma^2_\theta = s^2(T) - (1/k)\sum_{k=1}^{k} v_i$, where k = number of studies, and $s^2(T) = \sum_{k=1}^{k}\left[\left(T_i^2 - \overline{T}\right)^2/(k-1)\right]$, where $\overline{T}$ is the unweighted mean of T_1 through T_k (Shadish & Haddock, 1994, p. 274).

(Ho: $\beta_j = 0$), an approximate of χ^2 test (i.e., the sum of squares for model) was used with the degrees of freedom equal the number of predictors.

Tables 4A, 4B, and 4C show the results of regression analyses[2] under the mixed effects model. Type of CT and level of examinee age were significant moderators, with negative coefficients. This means that: first, the CBT ESs (CT minus P&P test scores) are relatively higher than the CAT ESs; second, the mean ESs scored by college or adult examinees are relatively lower than the mean ESs scored by other sample groups. When looking at CAT only (n=77), source and mathematics are significant moderators with positive coefficients. This means that the mean ESs reported in journals and the mean mathematics test ES are relatively higher than those of any other source and subject area, respectively, in CAT settings. When looking at CBT only (n=149), source of publication, level of examinee age, sample size and mathematics are significant moderators. The mean ESs reported in journals, the mean college students and adults' ESs and the mean mathematics ES are relatively lower than those of any other source, samples and subject area, respectively, in CBT settings also.

Synthesis II.

After removing nonindependent ESs by eliminating dependent effect sizes and creating composites, 146 ESs remain. The decision rules for eliminating studies were: first, remove all second trials if the same examinees took either or both modes twice; second, use the total score, if reported (the information about the intercorrelations is reported in the Table 5); third, use other research to impute the correlation(s) and compute composites if not reported. Twenty two ESs were removed due to the first 2 rules. Additionally, 73 ESs were combined into 15 composite ESs. Fifteen studies with more than one nonindependent ES were analyzed to see how different the composite ESs are through several methods of calculating composite ESs in Table 6.

With 146 effect sizes, the Q-between statistic of the homogeneity test results using composite effect sizes is 804.7 ($p \leq .000$, $df = 145$), which indicates heterogeneity. Fifty seven CAT ESs and eighty nine CBT ESs are also heterogeneous. This finding urges the author to use a random effect model rather than a fixed effect model for further analyses again. Table 7 summarizes the categorical analyses. The mean ES is −.001. The 95% confidence interval for d ranges from −.063 to

[2] Dummy variables are: Adaptive: CAT = 1, & CBT = 0; Journal: Journal = 1 & other sources = 0; College: college and adults = 1 & other samples = 0; Random: random with equivalence assignment = 1 & other designs = 0; Classroom: classroom test = 1, & other test types = 0; Mathematics: math = 1 & other subjects = 0 and English: English = 1 & other subjects = 0.

061, indicating that even though the CT scores are slightly higher than scores for P&P tests, the difference is not statistically significant. A significant Q statistic between adaptive types indicates that there is significant difference between CAT and CBT. While CAT has a negative mean ES, CBT has a positive mean ES. This means that examinees got higher scores for P&P tests than CAT, but lower scores for P&P tests than CBT. These results are the same as those from Synthesis I.

For CAT, while the Q statistics for source, samples, sample size, and test type are not significantly different ($p > .05$), publication year, content and design ($p < .05$) show significant differences. Some results from equivalence tests (based on 95% confidence interval) are different from the results with the typical method. The previously nonequivalent scores on sample size between 40 to 80 and English tests appear equivalent in this analysis. As a result, only sample size above 150 (for the sample size variable) and other subjects (e.g., science, medical knowledge, mechanical knowledge, education, etc.) show nonequivalence.

For CBT, only the Q statistic for publication year shows a significant difference ($p< .05$). The mean ESs for journal, military sample, sample size larger-than-150, standardized tests, English tests, other subjects tests, and nonrandom design, which were not equivalent in the typical method, were equivalent in this analysis. As result, the mean ESs for high school students, classroom tests, other cognitive tests and counter balanced design do not show nonequivalent scores.

Since the correlation between publication year and classroom was higher than .8 again, the publication year was again not used in the regression analyses. Tables 8A, 8B and 8C show the intercorrelations of moderators when using the G&O method. Type of CT is the only significant moderator variable (Table 8A). The negative coefficient means that the CBT ESs (CT minus P&P test scores) are relatively higher than the CAT ESs. For CAT only (n=57), source, sampling age, and mathematics are significant moderators (Table 8B). Journal and mathematics are significant moderators with positive coefficients. This means that the mean ESs reported in journals and the mean mathematics ES are relatively higher than those of any other source and subjects area respectively in CAT setting. The mean ESs scored by college or adult examinees are relatively lower than the mean ESs scored by other sample groups. For CBT only (n=89), there is no significant moderator variable (Table 8C).

Summary and Discussion

Nonindependence in meta-analysis

This synthesis had the goal of comparing potentially equivalent ability measures from computerized tests and paper-and-pencil tests while taking into account the nonindependence problem among effect sizes. Several researchers have pointed out that ignoring dependence between effect sizes underestimates the standard error and results in inflated Type I error (e.g., Chiu, 1997; Gleser & Olkin, 1994).

Table 9 summarizes the effect of adjusting for nonindependence on homogeneity tests. Two individual homogeneity tests for high school students and for sample sizes below 40 in CBT suggested homogeneity, using typical methods. But the studies appeared heterogeneous after avoiding nonindependence between effect sizes with G&O method. The rest of the individual homogeneity tests show the same results for two different methods (typical and G&O method). This result indicates that eliminating dependence between ESs does not affect the significance of homogeneity test too much (only 2 out of 50 individual homogeneity tests show different results).

Table 10 summarizes the comparison of the results of categorical analyses from the different methods. The Q statistics for source and test type in CAT, for source, test type and design in CBT, which were not significant with typical method, appeared significant after eliminating dependence of ESs. The ESs for CBT military sample, sample sizes greater than 150, and English tests and other subjects tests which were not equivalent with typical method, then appeared equivalent when dependence was eliminated. The opposite case happened for English tests in the CAT format, which appeared equivalent with typical meta-analysis methods, then were not equivalent in Synthesis II. This result can be explained by Figure 1. The two extreme mean ESs (−1.17 and −1.0) remained even after eliminating and combining dependent ESs, while the number of ESs were reduced from 20 to 12. Consequently, the mean ESs of English tests with G&O methods were reduced. Two extreme ESs affected the equivalence.

For CAT categories of sample size between 40 to 80, for CBT studies from journal, using standardized battery tests, and nonrandom designs were not found equivalent with the typical method, but then appeared equivalent with G&O. The 95% CI with lower absolute mean ESs has more chance to include zero in it, as the standard errors are the same.

Tables 11A, 11B and 11C show comparisons of regression analyses between typical and G&O approaches. One dominant comparison is the size of the standard error (s_j's). All of the

standard errors of Synthesis I were less than the standard errors of G&O method. Because of this, typical meta-analysis methods seem to be inflating the Type I error, especially for overall ESs (CAT and CBT combined) and for the CBT only regression analyses. However, this explanation does not hold when comparing the CAT regression analyses.

Overall, ignoring nonindependence between ESs tends to lead to underestimated standard errors and inflated Type I error rate when determining statistical significance tests. However, this is not always true because the means, dispersions, and distributions of ESs depend partly on the number of ESs, and partly on the methods adjusting for dependence of ESs.

Equivalence

The main findings for equivalence are shown in Table 7.

(1) On average, CTs are equivalent with P&P tests (overall ES equals −.001.).

(2) However, this equivalence is caused by combining the negative ES for CAT (−.147) and the positive ES for CBT (.097). Both of these ESs indicate statistically significant nonequivalence between both modes of CT and P&P tests.

(3) When the sample size is more than 150, the CAT scores are not equivalent with the P&P scores.

(4) CAT versions for mathematics and other cognitive measurements (recognition, logical reasoning, etc.) are equivalent with P&P versions, while CAT versions for English tests and other subjects tests (science, medical knowledge, mechanical knowledge, education, etc.) are not.

(5) CBT seems easier than the P&P version for high school students. This could be due to positive attitudes to CT or their excitement about taking CT.

(6) CB versions of classroom test are not equivalent with P&P versions, while standardized battery tests and author made CBTs are equivalent with the conventional tests.

(7) CBT versions for English tests, mathematics tests, and other subjects measurements are equivalent with the P&P tests, while CBT versions of other cognitive measurements are not.

Type of computerized is the most important variable when evaluating the equivalence between CT and P&P test (Table 11A). For CAT, mathematics, source and possibly the sampling age are significant variables (Table 11B). For CBT, the analyses did not find a significant moderator. These results imply that CB versions are relatively equivalent with the conventional tests, while CATs' equivalence is still affected by some moderators. However, one good situation is that

the most recent research (conducted between 1993 and 1996) have reported the equivalent mean effect sizes as those from the conventional tests (see Table 7).

These results are very different from the results of two previous meta-analyses especially those from Mead and Drasgow (1993) for two possible reasons. One reason is that Mead & Drasgow assumed that speededness is the most important matter in synthesizing scores from CT and P&P tests. However, as discussed before, speededness in CAT is not a factor; therefore it has been ignored in this synthesis. The type of computerized test in their synthesis was not a significant moderator, but it appeared as the most significant moderator in this synthesis. A second reason is the issue of nonindependence. Mead & Drasgow found 5 significant moderators among the 7 independent variables with a general regression model. This proportion went down to 1 of 8 independent variables when adjusting nonindependence between ESs which seems quite a bit lower even if we can not compare directly.

Limitations and Direction of Future Study

Meta-analysis is generally limited by the nature of the primary studies to which it is applied. This study synthesized 51 primary studies which include ability measures given as both P&P tests and either CAT or CBT. At least 14 unpublished technical reports which one previous meta-analysis synthesized were not included. Furthermore, another 20 studies were not included for this study because the studies did not satisfy the decision rules which were applied to literature retrieval. Thus, the results of this study may not generalize to all research in this area.

The author also has used own decision rules to adjust the nonindependence between ESs. Those rules also can not generalize to every single meta-analysis because other rules could be more appropriate for other syntheses. For instance, the author selected ES of the first trial when there were more than one trial (when the examinees took P&P test and CT both more than once). On the other hand, Kulik (1976) suggested that the results from only the most recent semester when an investigator reported data on the same course from several different semesters. Thus if a researcher uses his/her own decision rules to select more appropriate ES to adjust dependence, he/she could obtain results different from those of this study.

For the future research, three kinds of directions would be recommended. The first is including more specific moderators. For instance, one can include the speededness variable to analyze the ESs of the P&P tests and CBTs because it is a significant element of the equivalence between two modes as one previous meta-analysis concluded. Gender and anxiety are also potential

moderators, especially when considering the equity issues. But, the synthesis that will include these variables might have sufficient number of ESs.

Since 1989, investigations have appeared of the effect of self-adaptive testing (SAT) which seeks to minimize student anxiety and maximize student performance by allowing the examinee to have a chance to select items (Rocklin & O'Donnell, 1987). Several studies have compared SAT with CAT, finding that examinees receiving a self-adapted test obtained significantly higher mean proficiency estimates (Rocklin & O'Donnell, 1987; Wise, Plake, Johson, & Roos, 1992; Roos, Plake, & Wise, 1992; Vispoel & Coffman, 1992). Thus a meta-analysis for the self-adaptive testing will be needed to find either the equivalence between the P&P tests and SAT or the difference of ES between CAT and SAT in the near future.

Finally, several authors concluded that ignoring nonindependence between ES underestimates the standard error, and consequently inflates Type I error rate (e.g., Raudenbush, Becker, & Kalaian, 1988; Chiu, 1997 and so on). However, there has not been an empirical research that investigated how much affected by ignoring nonindependence the statistical power is. Thus, for example, a simulated statistical analysis is possible to show the power rates along with different number of ESs, different correlational coefficients between dependent ESs and/or different α levels

Tables

Table 1 Descriptive Information of Primary Studies

Author(s) (Pub. Year)	Test Name	Dependent Variables (Outcome measures)	Sample	Design	CAT
Baghi et al. (91)	MFTF	Math & Reading	High school	Counterbalanced	CAT
Baird & Silvern (92)	Author-made	Vocabulary	College	Random	CBT
Bejar & Weiss (78)	Class exam	Biology	College	CAT 1stP	CAT
Blackmore (86)	DAT	Verbal reasoning, Numerical ability, Abstract reasoning, Mechanical reasoning, Spatial reasoning, Language usage	G 12	Random	Both
Bugbee & Bernt (90)	Class exam	HS	College	Nonrandom	CBT
Chin et al. (91)	Science ach.	Science	G 10	Random	CBT
Dillon (92)	ObGyn	Medical science	Med. Sch.	Non- & Random	CBT
Dimock & Cormier (91)	DAT	Verbal reasoning	College	Counterbalanced	CBT
Eaves & Smith (86)	Class exam	Educational ability Media	College	Random	CBT
Eignor (93)	SAT	Verbal, Math	College	Random	CAT
English et al. (77)	Class exam	Ed. Measurement	College	Random	CAT
Federico (89)	Class exam	Recognition	Military	P&P 1st	CBT
Federico & Ligget (86)	Class exam	Recognition	Military	Counterbalanced	CBT
Glowacki et al (95)	Class exam	Ed. Computer Technology	College	Counterbalanced	CBT
Harrel et al (89)	MAB-Verbal	Information, Comprehension, Arithmetic reasoning, Similarities, Vocabulary, Verbal-IQ	College	Random	CBT
Henly et al (89)	DAT	Verbal reasoning, Numerical ability, Abstract reasoning, Mechanical reasoning, Spatial relations, Spelling, Language usage, Clerical speed & accuracy	College	Counterbalanced	CAT
Hoffman & Lundberg (76)	Class exam	Pharmacy knowledge	College	Matching	CBT
Horton & Lovitt (94)	Reading	Factual & Interpretive	High school	Counterbalanced	CBT
Huba (88)	WPT	Cognitive ability	Job Applicants	Counterbalanced	CBT
Kim & McLaen (95)	Class exam	Algebra	College	Random	CAT
Kovac (90)	Job related	Math, Vocabulary	Job Applicants	Random	CBT
Kuan (90)	Class exam	Computer knowledge	College	Nonrandom	Both
Lee & Hopkins (85)	AR	Arithmetic Reasoning	College	Random	CBT
Lee et al (86)	ASVAB	Arithmetic reasoning	Military	Random	CBT
Legg Buhr (90)	CLAST	Reading	College	Random	CAT
Legg & Buhr (92)	CLAST	Math, Reading, Writing	College	Nonrandom	CAT

Table 1 (Continued)

Author(s) (Pub. Year)	Test Name	Dependent Variables (Outcome measures)	Sample	Design	CAT
Legg & Buhr (87)	CLAST	Math	College	Nonrandom	CAT
Llabre et al. (87)	CMM	Verbal reasoning	College	Random	CBT
Lunz & Bergstrom (95)	BOR Certificate	Medical knowledge	Candidate	Random	CAT
Lunz & Bergstrom (91)	Med. Certificate	Medical knowledge	Medical Student	P&P 1st	CAT
MacLennan et al. (88)	MAB	Information, Comprehension, Arithmetic reasoning, Similarities, Vocabulary	College	Counterbalanced	CBT
Mazzeo et al. (91)	CLEP	Math, English	College	Counterbalanced	CBT
McDonald (88)	Class exam	Computation	G 3-6	Counterbalanced	CBT
Moon (92)	TOEFL	English	College	Random	CAT
Muhlestern (91)	CLEP	English Composition	College	Counterbalanced	CBT
Neal (91)	TASP	Reading, Math, Writing	High School	Counterbalanced	CBT
Olsen (90)	CAP	Math	G 3, 6	Counterbalanced	Both
Parshall & Kromrey (93)	GRE	Verbal, Analysis, Quantitative	College & up	Random	CBT
Peterson et al (96)	SPIQ	IQ	Patient	Counterbalanced	CBT
Power & Oneil (92)	Praxis	Math, Reading	College	CAT 1st	CBT
Rudy-Baese (79)	Class exam	Philosophy	College	Random	CBT
Russell & Haney (96)	NAEP, & Class	Language art, Science, Math, Writing	Middle school	Random	CBT
Sorensen (85)	KFRCT	Induction, Number Facility, General reasoning, Logical reasoning	Graduate	Counterbalanced	CBT
Spray et al (89)	Class exam	Radio repair	Military	Random	CBT
Tollefson (78)	Class exam	Ed. Measurement (Knowledge, Application)	Graduate	Random	CBT
Van de Vijver & Harseld (94)	GATB	Name comparison, Computation, Three-dimensional space, Vocabulary, Tool matching, Arithmetic reasoning, Form matching	Military	Matching	CBT
Vogel (94)	GRE	Verbal	College	Counterbalanced	CBT
Ward (94)	MSDE	Math	G 7, 8	P&P 1st	CAT
Ward et al. (89)	Class exam	Special Ed.	College	Random	CBT
Watkins & Kush (88)	Capitalization	Capitalization	Elementary	Random	CBT
Wise et al (1989)	Algebra	Algebra	College	Random	CBT

[1] MAFT: Maryland Functional Testing Program; DAT: Differential Aptitude Tests; ObGyn: Obstetrics & Gynecology by National Board of Medical Examiners; ASVAB: Armed Service Vocational Aptitude Battery; MAB: Multidimensional Aptitude Battery; WPT: Western Personnel Test; CLAST: College Level Academic Skills Test; CMM: California Short-Form Test of Mental Maturity; BOR Certificate: Board of Registry Medical Technologies Certificate; CLEP: College Level Examination Program; TOEFL: Test of English as a Foreign Language; TASP: Texas Academic Skills Program; CAP: California Assessment Program; SPIQ: Swedish Performance Intelligence Test; Praxis Series: Professional Assessments for Beginning Teachers; NAEP: National Assessment of Education Progress; KFRCT: Kit of Factor-Referenced Cognitive Test; GATB: General Aptitude Battery; MSDE: Maryland State Department of Education; PPST: paraprofessional Skills Test.

Table 2. Characteristics of Sample of Effect Sizes (n=226)

Characteristics	No. of studies (%)	No. of Effect sizes (%)
Type of computerized		
Computerized adaptive test	13 (25.0)	77 (34.1)
Computerized based test	35 (69.2)	149 (65.9)
Both	3 (5.8)	
Publication year		
1976 ~ 1979	5 (9.6)	19 (8.4)
1985 ~ 1988	12 (23.1)	67 (29.6)
1989 ~ 1992	23 (44.2)	93 (41.2)
1993 ~ 1996	11 (23.1)	47 (20.8)
Source		
Journals	25 (50.0)	106 (46.9)
Dissertation	11 (21.2)	57 (25.2)
Unpublished report	15 (28.8)	63 (27.9)
Sample		
Below high school	3 (7.7)	9 (4.0)
High school students	8 (15.4)	49 (21.7)
College students or up	35 (67.3)	147 (65.0)
Military	5 (9.6)	21 (9.3)
Sample size		
N =< 40		57 (25.2)
40 < N < 80		56 (24.8)
80 =< N < 150		57 (25.2)
150 =< N		56 (24.8)
Test type		
Standardized battery	31 (59.6)	163 (72.1)
Classroom exam	15 (28.8)	52 (23.0)
Author made	4 (9.6)	11 (4.8)
Battery & Classroom exam both	1 (1.9)	
Test Content		
English		71 (31.4)
Mathematics		48 (21.2)
Other subjects (Science, Education, Mechanic, Medical, etc.)		57 (25.2)
Others general cognitive abilities (IQ, recognition, etc.)		50 (22.1)
Design		
Random	35 (69.2)	79 (35.0)
Nonrandom	4 (7.7)	15 (6.6)
P&P 1st		32 (14.2)
CAT 1st		46 (20.4)
Counter balanced	12 (23.1)	54 (23.9)

Table 3. Results of Categorical Analysis When Using Typical Meta-Analysis with Random Effects Model (n=226)

Variables	df	Q	p	T	Variance	SE	95% CI
Total	225	241.5	.215	.019	.0006	.025	-.030 ~ .068
Type of computerized (bet.)	1	19.3	.000				
Within groups	224	222.2					
CAT	76	60.9	.896	-.125	.0017	.041	-.206 ~ -.044
CBT	148	161.3	.215	.103	.0010	.031	.041 ~ .164
CAT							
Publication year (bet.)	3	22.8	.000				
Within groups	73	90.3					
1976 ~ 1979	7	11.9	.105	-.517	.0089	.094	-.702 ~ -.332
1985 ~ 1988	8	3.1	.931	-.051	.0153	.124	-.294 ~ .192
1989 ~ 1992	39	71.7	.001	-.126	.0017	.042	-.208 ~ -.044
1993 ~ 1996	19	3.7	.999	.011	.0036	.060	-.106 ~ .129
Source (bet.)	2	13.4	.001				
Within groups	74	99.7					
Journal	20	3.47	.999	.018	.0027	.052	-.085 ~ .121
Dissertation	24	18.0	.805	-.134	.0047	.069	-.269 ~ .000
Report	30	78.2	.000	-.240	.0022	.047	-.332 ~ -.148
Sample (bet.)	1	1.0	.315				
Within groups	74	110.1					
High school	24	55.4	.000	-.099	.0024	.049	-.195 ~ -.004
College & up	49	54.7	.267	-.164	.0018	.042	-.246 ~ -.082
(< High school)	1						
Sample size (bet.)	3	1.84	.605				
Within groups	75	111.3					
=< 40	10	3.8	.957	-.055	.0180	.134	-.318 ~ .208
40 < N < 80	20	12.6	.896	-.172	.0057	.075	-.319 ~ -.024
80 =< N < 150	9	1.0	.999	-.035	.0076	.087	-.206 ~ .135
150 =<	34	93.9	.000	-.139	.0015	.039	-.215 ~ -.063
Test type (bet.)	1	11.0	.001				
Within groups	75	99.3					
Classroom exam	9	22.1	.009	-.406	.0079	.089	-.580 ~ -.232
Standardized battery	64	77.2	.124	-.091	.0012	.034	-.158 ~ -.023
(Author made)	1						
Content (bet.)	3	11.9	.008				
Within groups	73	101.2					
English	19	50.6	.000	-.137	.0032	.056	-.247 ~ -.027
Math	17	4.1	.999	.007	.0034	.059	-.108 ~ .122
Other subjects	28	37.8	.102	-.272	.0034	.058	-.386 ~ -.158
Other Cognitive	9	8.7	.470	-.070	.0080	.089	-.245 ~ .105
Design (bet.)	3	20.2	.000				
Within groups	73	87.2					
Random	15	9.9	.827	-.054	.0056	.075	-.201 ~ .093
P&P 1st	12	23.6	.023	-.140	.0061	.078	-.294 ~ .013
CAT 1st	28	50.1	.006	-.316	.0031	.056	-.426 ~ -.206
Counter balanced	16	3.5	.999	.032	.0033	.057	-.079 ~ .144
(Nonrandom)	1						

Table 3 (Continued)

Variables	df	Q	p	T	Variance	SE	95% CI	
CBT								
Publication year (bet.)	3	21.7	.000					
Within groups	14	129.4						
1976 ~ 1979	10	9.8	.459	.274	.0144	.120	.038 ~	.509
1985 ~ 1988	57	27.4	.999	-.018	.0031	.056	-.128 ~	.091
1989 ~ 1992	52	64.1	.121	.036	.0028	.052	-.067 ~	.139
1993 ~ 1996	26	28.1	.354	.365	.0052	.072	.222 ~	.505
Source (bet.)	2	7.8	.020					
Within groups	14	143.2						
Journal	84	52.7	.997	.095	.0018	.042	.012 ~	.178
Dissertation	31	45.3	.046	-.049	.0059	.077	-.199 ~	.102
Report	31	45.2	.048	.246	.0045	.067	.104 ~	.368
Sample (bet.)	3	13.0	.005					
Within groups	14	138.0						
High school	23	16.9	.813	.306	.0070	.084	.142 ~	.470
College & up	96	101.8	.324	.027	.0017	.041	-.053 ~	.107
Military	20	18.2	.575	.238	.0063	.079	.083 ~	.393
< High school	6	1.1	.980	-.026	.0216	.147	-.314 ~	.262
Sample size (bet.)	3	3.4	.332					
Within groups	14	147.7						
=< 40	45	37.3	.787	.022	.0047	.069	-.112 ~	.157
40 < N < 80	34	15.8	.997	.114	.0049	.070	-.023 ~	.251
80 =< N < 150	46	80.0	.001	.088	.0029	.054	-.018 ~	.194
150 =<	20	14.6	.798	.204	.0053	.073	.061 ~	.347
Test type (bet.)	2	11.4	.003					
Within groups	14	139.7						
Classroom exam	41	35.0	.733	.168	.0036	.060	.051 ~	.285
Standardized battery	97	75.1	.951	.108	.0016	.040	.030 ~	.187
Author made	8	29.5	.000	-.353	.0204	.143	-.632 ~	-.073
Content (bet.)	3	7.6	.055					
Within groups	14	143.5						
English	50	45.0	.674	.128	.0032	.056	.017 ~	.239
Math	29	38.1	.120	-.068	.0050	.071	-.206 ~	.071
Other subjects	27	31.1	.265	.147	.0054	.073	.003 ~	.291
Other Cognitive	39	29.2	.872	.172	.0040	.063	.047 ~	.296
Design (bet.)	4	13.0	.011					
Within groups	14	138.1						
Random	62	74.2	.138	.089	.0026	.051	-.012 ~	.189
P&P 1st	18	18.8	.401	-.109	.0079	.089	-.283 ~	.064
CBT 1st	16	19.8	.228	.049	.0091	.095	-.138 ~	.236
Counter balanced	36	16.7	.997	.156	.0045	.067	.024 ~	.287
Nonrandom	12	8.5	.748	.349	.0094	.097	.155 ~	.535

Table 4A. Regression Analysis for All When Using Typical Meta-Analysis with Mixed Effects Model (n=226)

Variables	B	Beta	SE	S	z
Intercept	.3647				
Adaptive	−.2965	−.3690	.0503	.0479	6.190**
Journal	−.0820	−.0976	.0602	.0573	1.431
College students	−.1614	−.1866	.0629	.0599	2.694**
Random	−.0223	−.0269	.0575	.0520	0.429
Classroom test	−.0401	−.0437	.0660	.0628	0.639
Sample size	.0001	.0306	.0001	.0001	1.000
Math	−.1282	−.1363	.0770	.0733	1.749
English	−.0782	−.0976	.0642	.0611	1.280

χ^2_8 (model significance) = 31.07** MSE = 1.1042

** p < .01

Table 4B. Regression Analysis for CAT When Using Typical Meta-Analysis with Mixed Effects Model (n=77)

Variables	B	Beta	SE	S	z
Intercept	−.4506				
Journal	.3287	.4900	.0893	.0805	4.083**
College students	.1363	.2038	.0917	.0826	1.537
Random	−.0124	−.0143	.0936	.0843	0.1471
Classroom test	−.2004	−.1990	.1250	.1126	1.780
Sample size	−.0001	−.0629	.0002	.0002	0.500
Math	.3008	.4081	.0999	.0900	3.342**
English	.1029	.1427	.0938	.0845	1.217

χ^2_8 (model significance) = 35.64** MSE = 1.2327

** p < .01

Table 4C. Regression Analysis for CBT When Using Typical Meta-Analysis with Mixed Effects Model (n=149)

Variables	B	Beta	SE	S	z
Intercept	.6493				
Journal	−.2518	−.2680	.0750	.0746	3.375**
College students	−.3003	−.3003	.0861	.0856	3.508**
Random	−.0445	−.0548	.0688	.0684	0.651
Classroom test	−.0402	−.0518	.0773	.0768	0.523
Sample size	.0004	.1823	.0002	.0002	2.000*
Math	−.4042	−.3943	.0995	.0989	4.087**
English	−.1493	−.1853	.0776	.0771	1.936

χ^2_8 (model significance) = 33.43** MSE = 1.0115

** p < .01, * p < .05

Table 5. Information of Correlation(s)

Study	Information on correlation(s)	Test
Blackmore (86)	Henly et al. (89)	DAT [1]
Harrell et al. (87)	Wallbrown et al. (88)[2]	MAB [3]
Henly et al. (89)	correlation reported	DAT
Kovac (89)	Heyn & Hilton (82) [4]	Math, Vocabulary
Legg & Buhr (92)	Neal (91)	CLAST [5]
Neal (91)	correlations reported	TASP [6]
Russell & Haney (96)	correlation reported	NAEP [7]
Sorensen (85)	0.5	KFRCT [8]
Viver & Harsvel (94)	Correlations reported	GATB [9]

1) DAT: Differential Aptitude Tests

2) Wallbrown, F.H., Carmin, C.N., & Barnett, R.W. (1988). Psychological Reports, 62, 871-878.

3) MAB: Multidimensional Aptitude Battery

4) Heyns, B., & Hilton, T. L. (1982). The cognitive tests for high school and beyond: An assessment. Sociology of education, 55, 89-102.

5) CLAST: College Level Academic Skills Test

6) TASP: Texas Academic Skills Program

7) NAEP: National Assessment of Education Progress

8) KFRCT: Kit of Factor-Referenced Cognitive Test. Rosenthal & Rubin (1988) recommend of .5 correlational coefficient for cognitive measures.

9) GATB: General Aptitude Test Battery

Table 6. Unbiased Composite Effect Sizes Computed by Typical and G&O Approaches from Fifteen Studies

Study	Groups	nc[1]	np[2]	No of d's	$\bar{d}$[3]	G&O's δ[4]	r of d(s)[5]
Russell & Haney (96)	Middle school	48	72	3	.71	.59	.82
Sorensen (85)	Male	20	13	4	−.07	.10	.50
	Female	29	34	4	.14	.20	.50
Henley et al. (89)	Sample A	171	171	8	.03	−.04	.60
	Sample B	161	161	8	.03	−.04	.55
Harrell et al. (87)	1st trial	20	20	6	−.01	−.05	.68
	P&P 1st	20	20	6	−.53	−.55	.68
	CAT 1st	20	20	6	.04	−.29	.68
Legg & Buhr (92)	College	518	518	3	.25	.24	.39
Blackmore (86)	CAT	24	24	6	−.15	−.03	.56
	CBT	24	24	6	.06	.12	.56
Kovac (89)	Job applicants	59	62	2	−1.37	−1.37	.60
Neal (91)	Male	20	20	3	.18	.22	.31
	Female	15	15	3	.31	.59	.31
Vijver & Harsvel (94)	Military	163	163	7	.41	.25	.28

1) Number of subjects who took CAT.
2) Number of subjects who took P&P.
3) Average unweighted effect size
4) Gleser & Olkin's composite effect size
5) Average intercorrelation among effect sizes

Table 7. Result of Categorical Analysis When Using G&O Method with Random Effects Model (n=146)

Variables	df	Q	p	T	Variance	SE	95% CI
Total	145	155.3	.264	-.001	.0010	.031	-.063 ~ .061
Type of computerized (bet.)	1	14.5	.000				
Within groups	145	140.9					
CAT	56	51.7	.636	-.147	.0025	.050	-.244 ~ -.050
CBT	88	89.1	.446	.097	.0017	.041	.017 ~ .177
CAT							
Publication year (bet.)	3	19.5	.000				
Within groups	53	62.7					
1976 ~ 1979	4	8.2	.084	-.568	.0183	.135	-.832 ~ -.303
1985 ~ 1988	3	.4	.941	.150	.0431	.208	-.256 ~ .557
1989 ~ 1992	27	50.9	.004	-.227	.0034	.058	-.342 ~ -.113
1993 ~ 1996	19	3.2	.999	.009	.0042	.065	-.118 ~ .135
Source (bet.)	2	3.6	.162				
Within groups	54	79.5					
Journal	7	1.0	.995	-.005	.0095	.098	-.186 ~ .196
Dissertation	19	15.2	.713	-.133	.0065	.081	-.291 ~ .026
Report	28	62.4	.002	-.205	.0028	.053	-.310 ~ -.101
Sample (bet.)	1	1.6	.201				
Within groups	53	79.3					
High school	7	35.8	.000	-.267	.0080	.089	-.442 ~ -.092
College & up (< High school)	46	42.5	.621	-.138	.0022	.047	-.230 ~ -.046
Sample size (bet.)	3	2.8	.430				
Within groups	146	79.4					
=< 40	10	3.6	.964	-.056	.0191	.138	-.327 ~ .214
40 < N < 80	14	10.4	.730	-.181	.0086	.093	-.362 ~ .000
80 =< N < 150	9	.9	.999	-.036	.0087	.093	-.218 ~ .147
150 =<	20	64.5	.000	-.196	.0031	.055	-.304 ~ -.087
Test type (bet.)	1	3.5	.063				
Within groups	53	75.9					
Classroom exam	6	17.6	.007	-.369	.0146	.121	-.606 ~ -.132
Standardize battery (Author made)	47	58.3	.125	-.130	.0020	.045	-.217 ~ -.042
Content (bet.)	3	11.3	.010				
Within groups	53	70.8					
English	11	34.9	.000	-.242	.0063	.080	-.398 ~ -.086
Math	14	2.9	.999	.023	.0048	.070	-.114 ~ .159
Other subjects	22	30.0	.212	-.293	.0056	.075	-.440 ~ -.146
Other Cognitive	6	6.0	.424	-.095	.0146	.121	-.332 ~ .142
Design (bet.)	3	9.7	.022				
Within groups	52	69.9					
Random	10	7.1	.712	-.018	.0084	.091	-.198 ~ .161
P&P 1st	12	20.1	.064	-.134	.0070	.084	-.298 ~ .030
CAT 1st	26	41.5	.028	-.285	.0040	.063	-.410 ~ -.161
Counter balanced (Nonrandom)	4	1.1	.895	.050	.0133	.115	-.176 ~ .276

Table 7 (Continued)

Variables	df	Q	p	T	Variance	SE	95% CI
CBT							
Publication year (bet.)	3	8.5	.037				
Within groups	86	78.8					
1976 ~ 1979	5	6.9	.225	.459	.0320	.179	.108 ~ .809
1985 ~ 1988	20	10.7	.954	-.008	.0080	.090	-.184 ~ .168
1989 ~ 1992	44	43.3	.500	.052	.0031	.056	-.058 ~ .162
1993 ~ 1996	16	17.8	.335	.240	.0087	.093	.057 ~ .423
Source (bet.)	2	2.5	.289				
Within groups	87	84.8					
Journal	44	22.8	.997	.094	.0033	.057	-.018 ~ .206
Dissertation	15	24.6	.056	-.039	.0115	.107	-.249 ~ .171
Report	27	37.3	.089	.163	.0051	.071	.024 ~ .303
Sample (bet.)	3	3.5	.322				
Within groups	86	83.7					
High school	10	3.3	.973	.276	.0142	.119	.042 ~ .510
College & up	55	67.9	.114	.066	.0028	.052	-.036 ~ .169
Military	14	11.4	.657	.138	.0091	.096	-.049 ~ .326
< High school	6	1.2	.979	-.025	.0211	.145	-.310 ~ .259
Sample size (bet.)	3	.784	.853				
Within groups	146	86.5					
=< 40	17	16.4	.499	.145	.0128	.113	-.077 ~ .366
40 < N < 80	25	14.1	.961	.138	.0064	.080	-.019 ~ .295
80 =< N < 150	31	51.2	.013	.059	.0042	.065	-.068 ~ .186
150 =<	12	4.8	.964	.090	.0084	.092	-.090 ~ .270
Test type (bet.)	2	3.9	.277				
Within groups	87	83.4					
Classroom exam	34	30.0	.664	.143	.0043	.066	.015 ~ .272
Standardized	45	32.0	.927	.102	.0032	.056	-.009 ~ .212
Author made	7	21.3	.003	-.183	.0233	.153	-.482 ~ .116
Content (bet.)	3	2.0	.571				
Within groups	86	85.2					
English	26	18.7	.848	.084	.0058	.076	-.065 ~ .233
Math	15	8.8	.887	.005	.0089	.094	-.180 ~ .190
Other subjects	25	45.9	.007	.103	.0058	.076	-.046 ~ .253
Other Cognitive	19	11.8	.895	.185	.0075	.087	.015 ~ .355
Design (bet.)	4	7.4	.114				
Within groups	85	79.8					
Random	32	16.3	.132	.030	.0114	.107	-.179 ~ .240
P&P 1st	11	38.6	.197	.057	.0047	.068	-.077 ~ .192
CAT 1st	11	10.4	.495	-.075	.0117	.108	-.287 ~ .137
Counter balanced	24	10.9	.990	.234	.0067	.082	.074 ~ .394
Nonrandom	6	3.7	.720	.256	.0183	.135	-.009 ~ .520

Table 8A. Regression Analysis for All When Using G&O Method with Mixed Effects Model (n=146)

Variables	B	Beta	SE	S	z
Intercept	.1474				
Adaptive	-.2399	-.3002	.0728	.0696	3.447**
Journal	-.0101	-.0227	.0748	.0715	0.141
College students	-.0290	-.0303	.0877	.0838	0.346
Random	.0033	.0039	.0779	.0744	0.044
Classroom test	.0106	.0122	.0834	.0797	0.133
Sample size	-.0001	-.0743	.0001	.0001	1.000
Math	.0436	.0478	.0954	.0950	0.459
English	-.0470	-.0339	.0864	.0826	0.569

χ^2_8 (model significance) = 17.61* MSE = 1.0953

** p < .01, * p < .05

Table 8B. Regression Analysis for CAT When Using G&O Method with Mixed Effects Model (n=57)

Variables	B	Beta	SE	S	z
Intercept	-.5864				
Journal	.3036	.4021	.1132	.0961	3.159**
College students	.2092	.2500	.1207	.1024	2.043*
Random	.1080	.1171	.1227	.1041	1.037
Classroom test	-.1466	-.1245	.1621	.1376	1.065
Sample size	-.0001	-.1020	.0002	.0002	0.500
Math	.4424	.5657	.1305	.1108	3.993**
English	.1253	.1477	.1358	.1153	1.087

χ^2 - (model statistic) = 33.22** MSE = 1.3884

** p < .01, * p < .05

Table 8C. Regression Analysis for CBT When Using G&O Method with Mixed Effects Model (n=89)

Variables	B	Beta	SE	S	z
Intercept	.4711				
Journal	-.1619	-.1986	.1000	.1005	1.611
College students	-.1018	-.2097	.1180	.1186	0.858
Random	-.0619	-.0773	.1026	.1031	0.600
Classroom test	.0029	.0036	.0982	.0987	0.029
Sample size	-.00003	-.0117	.0002	.0002	0.150
Math	-.2218	-.2264	.1303	.1310	1.693
English	-.1210	-.1451	.1154	.1160	1.043

χ^2- (model significance) = 6.73 MSE = .9897

Table 9. Comparison of Results of Homogeneity Tests Between Typical and G&O methods

	df	Typical	df	G&O	df	Typical	df	G&O
Total	225	Hete.	145	Hete.				
		<CAT>				<CBT>		
	76	Hete.	56	Hete	148	Hete.	88	Hete.
Publication Year								
1976 ~ 1979	7	Hete.	4	Hete.	10	Hete.	5	Hete.
1985 ~ 1988	8	Homo.	3	Homo.	57	Homo.	20	Homo.
1989 ~ 1992	39	Hete.	27	Hete.	52	Hete.	44	Hete.
1993 ~ 1996	19	Homo.	19	Homo.	26	Hete.	16	Hete.
Source								
Journal	20	Homo.	7	Homo.	84	Hete.	44	Hete.
Dissertation	24	Hete.	19	Hete.	31	Hete.	15	Hete.
Report	30	Hete.	28	Hete.	31	Hete.	27	Hete.
Sample								
High school	24	Hete.	7	Hete.	23	Hete.	10	Homo.
College & up	49	Hete.	46	Hete.	96	Hete.	55	Hete.
Military					20	Hete	14	Hete.
< High school	1		1		6	Homo.	6	Homo.
Sample size								
=< 40	11	Homo.	10	Homo.	45	Hete.	17	Homo.
40 < N < 80	21	Homo.	14	Homo.	34	Homo.	25	Homo.
80 =< N < 150	10	Homo.	9	Homo.	46	Hete.	31	Hete.
150 =<	34	Hete.	20	Hete.	20	Hete.	12	Hete.
Test type								
Classroom exam	9	Hete.	6	Hete.	41	Hete.	34	Hete.
Standardized battery	64	Hete.	47	Hete.	97	Hete.	45	Hete.
Author made	1				8	Hete.	7	Hete.
Test content								
English	19	Hete.	11	Hete.	50	Hete.	26	Hete.
Math	17	Homo.	14	Homo.	29	Hete.	15	Hete.
Other subjects	28	Hete.	22	Hete.	27	Hete.	25	Hete.
Other Cognitive	9	Hete.	6	Hete.	39	Hete.	19	Hete.
Design								
Random	15	Hete.	10	Hete.	62	Hete.	32	Hete.
P&P 1st	12	Hete.	12	Hete.	18	Hete.	11	Hete.
CAT 1st	28	Hete.	26	Hete.	16	Hete.	11	Hete.
Counter balanced	18	Homo.	4	Homo.	36	Homo.	24	Homo.
Nonrandom	1				12	Hete.	6	Hete.

Table 10 Comparison of Results of Categorical Analyses Between Typical and G&O Approaches with Random Effects Model

Variables	Typical			G&O			Typical			G&O		
	df	Homo.	T	df	Homo.	T	df	Homo.	T	df	Homo	T
Total	225	Homo.	.019	145	Homo.	-.001						
Adaptive type (bet.)	1	Hete.		1	Hete.							
Within groups	224			144								
CAT	76	Homo.	-.125*	56	Homo.	-.147*						
CBT	148	Homo.	.103*	88	Homo.	.097*						
			<CAT>						<CBT>			
Pubyear (bet.)	3	Hete.		3	Hete.		3	Hete.		3	Hete.	
Within groups	73			53			145			86		
1976 - 1979	7	Homo.	-.517*	4	Homo.	-.568*	10	Homo.	.274*	5	Homo	.459*
1985 - 1988	8	Homo.	-.051	3	Homo.	.150	57	Homo.	-.018	20	Homo	-.008
1989 - 1992	39	Hete.	-.126*	27	Hete.	-.227*	52	Homo.	.036	44	Homo	.052
1993 - 1996	19	Homo.	.001	19	Homo.	.009	26	Homo.	.365*	16	Homo	.240*
Source (bet.)	2	Hete.		2	Homo.		2	Hete.		2	Homo	
Within groups	74			54			147			86		
Journal	20	Homo.	.018	7	Homo.	-.005	84	Homo.	.095*	44	Homo	.094
Dissertation	24	Homo.	-.134	19	Homo.	-.133	31	Hete.	-.049	15	Homo	-.039
Report	30	Hete.	-.240*	28	Hete.	-.205*	31	Hete.	.246*	27	Homo	.163*
Sample (bet.)	1	Homo.		1	Homo.		3	Hete.		3	Homo	
Within groups	74			53			145			85		
High school	24	Hete.	-.099*	7	Hete.	-.267*	23	Homo.	.306*	10	Homo	.276*
College & up	49	Homo.	-.164*	46	Homo.	-.138*	96	Homo.	.027	55	Homo	.066
Military							20	Homo.	.238*	14	Homo	.138
<High School	1			1			6	Homo.	-.026	6	Homo	-.025
Sample size (bet.)	3	Homo.		3	Homo.		3	Homo.		3	Homo	
Within groups	75			146			146			146		
=< 40	10	Homo.	-.055	10	Homo.	-.056	45	Homo.	.022	17	Homo	.145
40 < N < 80	20	Homo.	-.172*	14	Homo.	-.181	34	Homo.	.114	25	Homo	.138
80 =< N < 150	9	Homo.	-.035	9	Homo.	-.036	46	Hete.	.088	31	Hete.	.059
150 =<	36	Hete.	-.139*	20	Hete.	-.196*	20	Homo.	.204*	12	Homo	.090
Test type (bet.)	1	Hete.		1	Homo.		2	Hete.		2	Homo	
Within groups	75			53			147			86		
Classroom exam	9	Hete.	-.406*	6	Hete.	-.369*	41	Homo.	.168*	34	Homo	.143*
Standardized	64	Homo.	-.091*	47	Homo.	-.130*	97	Homo.	.108*	45	Homo	.102
Author made							8	Hete.	-.353	7	Hete.	-.183
Content (bet.)	3	Hete.		3	Hete.		3	Homo.		3	Homo	
Within groups	73			53			146			86		
English	19	Hete.	-.137*	11	Hete.	-.242*	50	Homo.	.128*	26	Homo	.057
Math	17	Homo.	.007	14	Homo.	.023	29	Homo.	-.068	15	Homo	.005
Other subjects	28	Homo.	-.272*	22	Homo.	-.293*	27	Homo.	.147*	25	Hete.	.103
Other Cognitive	9	Homo.	-.070	6	Homo.	-.095	39	Homo.	.172*	19	Homo	.185*
Design (bet.)	3	Hete.		3	Hete.		4	Hete.		4	Homo	
Within groups	73			52			144			85		
Random	15	Homo.	-.054	10	Homo.	-.018	62	Homo.	.049	32	Homo	.030
P&P 1st	12	Hete.	-.140	12	Homo.	-.134	18	Homo.	.089	11	Homo	.057
CAT 1st	28	Hete.	-.316*	26	Hete.	-.285*	16	Homo.	-.109	11	Homo	-.075
Counter balanced	16	Homo.	.032	4	Homo.	.050	36	Homo.	.156*	24	Homo	.234*
Nonrandom							12	Homo.	.349*	6	Homo	.256

BEST COPY AVAILABLE

Table 11A. Comparison of Regression Analysis Between Typical and G&O Methods for All with Mixed Effects Model

	Typical		G&O	
Variables	Beta	SE	Beta	SE
Adaptive	-.3690**	.0479	-.3002**	.0696
Journal	-.0976	.0573	-.0227	.0715
College students	-.1866**	.0599	-.0303	.0838
Random	-.0269	.0520	.0039	.0744
Classroom test	-.0437	.0628	.0122	.0797
Sample size	.0306	.0001	-.0743	.0001
Math	-.1363	.0733	.0478	.0950
English	-.0976	.0611	-.0339	.0826
	χ^2_8 =	31.07**	χ^2_8 =	17.61*

** p < .01, * p < .05

Table 11B. Comparison of Regression Analysis Between Typical and G&O Methods for CAT with Mixed Effects Model

	Typical		G&O	
Variables	Beta	SE	Beta	SE
Journal	.4900**	.0805	.4021**	.0961
College students	.2038	.0826	.2500*	.1024
Random	-.0143	.0843	.1171	.1041
Classroom test	-.1990	.1126	-.1245	.1376
Sample size	-.0629	.0002	-.1020	.0002
Math	.4081**	.0900	.5657**	.1108
English	.1427	.0845	.1477	.1153
	χ^2_8 =	35.64**	χ^2_8 =	33.22**

** p < .01, * p < .05

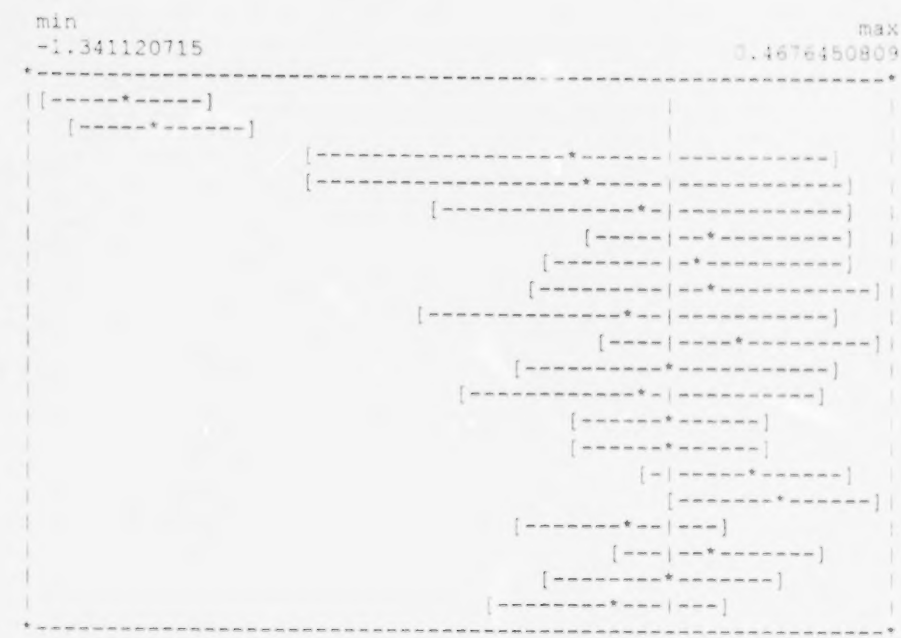
Table 11C. Comparison of Regression Analysis Between Typical and G&O Methods for CBT with Mixed Effects Model

	Typical		G&O	
Variables	Beta	SE	Beta	SE
Journal	-.2680**	.0746	-.1986	.1005
College students	-.3003**	.0856	-.2097	.1186
Random	-.0548	.0684	-.0773	.1031
Classroom test	-.0518	.0768	.0036	.0987
Sample size	.1823*	.0002	-.0117	.0002
Math	-.3943**	.0989	-.2264	.1310
English	-.1853	.0771	-.1451	.1160
	χ^2_8 =	33.43**	χ^2_8 =	6.73

** p < .01, * p < .05

(1) Typical method

LLIM	ULIM	T
-1.34	-1.00	-1.17
-1.29	-0.91	-1.10
-0.78	0.35	-0.22
-0.76	0.37	-0.20
-0.50	0.38	-0.06
-0.19	0.37	0.09
-0.28	0.38	0.05
-0.29	0.43	0.07
-0.55	0.36	-0.10
-0.17	0.47	0.15
-0.35	0.35	0.00
-0.45	0.31	-0.07
-0.21	0.21	0.00
-0.23	0.21	-0.01
-0.05	0.37	0.16
0.00	0.44	0.22
-0.32	0.10	-0.11
-0.13	0.31	0.09
-0.26	0,22	-0.02
-0.39	0.11	-0.14



(2) G&O method

LLIM	ULIM	T
-1.34	-1.00	-1.17
-1.29	-0.91	-1.10
-0.50	0.38	-0.06
-0.19	0.37	0.09
-0.28	0.38	0.05
-0.29	0.43	0.07
-0.55	0.36	-0.10
-0.17	0.47	0.15
-0.35	0.35	0.00
-0.45	0.31	-0.07
-0.26	0.22	-0.02
-0.39	0.11	-0.14

Figure 1. Comparison of Distributions of English Tests for CAT

References

Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin, 99(3)*, 388-399.

Bergstrom, B.A.(1992). *Ability measure equivalence of computer adaptive and pencil and paper test: a research synthesis*. Paper presented at the Annual American Educational Research Association (San Francisco, CA, April 20-24, 1992).

Birenbaum, M., & Tatsuoka, K.K. (1987). Effects of "on line" test feedback on the seriousness of subsequent errors. *Journal of Educational Measurement, 24(2)*, 145-155.

Bond, L. (1989). The effects of special preparation on measures of scholastic ability. In R.L. Linn (Ed.). *Educational measurement (3rd. Ed.)*, 429-444.

Brown, J.M., & Weiss, D.J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In L. Linn (ed.), Educational Measurement (3rd ed., pp. 367-408). New York: Mcmillan.

Chiu, C.W.T. (1997). *Synthesizing multiple outcome measures with missing data: A sensitivity analysis on the effect of metacognitive reading*. Unpublished apprenticeship paper, CEPSE Department, MSU.

Cole, D.A., Maxwell, S.E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115 (3)*, 465-474.

Cooley, W.W., & Lohnes, P.R. (1976). *Evaluation research in education*. Irvington Publishers, New York.

Cooper, H. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality an Social Psychology, 37*, 131-146.

Cooper, H. & Hedges, L. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage.

English, R.A., Reckase, M.D., & Patience, W.M. (1977). Application of tailored testing to achievement measurement. Behavior Research Methods and Instrumentation, 9, 158-161.

Federico, P. (1989). *Computer-based and paper-based measurement of recognition performance*. Navy Personnel Research and Development Center (NPRDC-TR-89-7).

Federico, P., & Liggett, N.L. (1989). *Computer-based and paper-based measurement of semantic knowledge*. Navy Personnel Research and Development Center, NPRDC-TR-89-4.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. Hedges (Eds.) *The handbook of research synthesis (pp. 339-355)*. New York: Russell Sage.

Greaud, V.A. & Green, B.F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23-34.

Green, B.F. (1987). Construct validity of computer-based tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. & Reckase, M.D. (1984). Technical guideline for assessing computerized adaptive tests. *Journal of Educational Measurement, 21(4)*, 347-360.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory*. Vancouver: Educational Research Institute of British Columbia.

Haney, W. (1991). We must take care: Fitting assessments to function. In V. Perrone (Ed.), *Expanding student assessment* (pp. 47-71). Alexandria, VA: Association for Supervision and Curriculum Development.

Hansen, D.N. (1969). An investigation of computer-based science testing. In R.C. Atkinson & H.A. Wilson (Eds.) *Computer-assisted instruction: A book of readings* (pp. 209-226). New York: Academic.

Hedges, L.V. (1994). Fixed effects models. In H. Cooper & L. Hedges (Eds.) *The handbook of research synthesis (pp. 285-299)*. New York: Russell Sage.

Heissen, R.K., & Glass, C.R., & Knight, L.A. (1987). Assessing computer anxiety: Development and validation of the computer Anxiety Rating Scale. *Computers in Human Behavior, 3*, 49-59.

Holmes, C.T. & Matthews, K.M. (1984). The effect of nonpromotion on elementary and junior school pupils: a meta-analysis. *Review of Educational Research, 54*, 225-236.

Iaffaldano, M.T., & Muchinsky, P.M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97*, 251-273.

Kovac, R.M.N. (1990). *The effects of computerized selection tests on job applicant performance*. Unpublished dissertation. DePaul University.

Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics, 8(2)*, 93-101.

Kuan, T.H. (1991). *A comparison of paper-and-pencil, computer-administered, computerized feedback, and computerized adaptive testing*. Unpublished dissertation. Mississippi State University.

Kulik, J.A. (1976). PSI: A formative evaluation. In B.A. Green, Jr. (Ed.), *Personalized instruction in higher education: Proceedings of the second national conference*. Washington, D.C.: Center for Personalized Instruction.

Kulik, J.A., Kulik, C.C. & Cohen, P.A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34(4)*, 307-318.

Landman, J.T., & Dawes, R.M. (19982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist, 37(5)*, 504-516.

Lee, J.A. (1986). The effects of past computer experience on computerized aptitude test performance. *Educational and Psychological Measurement, 46*, 727-733.

Llabre, M. M., Clments, N. E., Fitzhugh, K. B., Lancelotta, G., Mazzagatti, R. D., & Quinones, N. (1987). The effect of computer-administered testing on test anxiety and performance. *Journal of Educational Computing Research, 3(4)*, 429-433.

Lockheed, M.E. (1985). Women, girls, and computers: A first look at the evidence. *Sex Roles, 13*, 115-122.

Lunz, M., Bergstrom, B., & Wright, B. D. (1992). He effect of review on student ability and test efficiency for computer adaptive tests. *Applied Psychological Measurement, 16*, 33-40.

Martinez, M.E., & Mead, N.A. (1988). *Computer competence: The first national assessment*. Princeton, NJ: Educational Testing Service.

Maurelli, V.A., & Weiss, D.J. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries* (Research Rep. No. 81-4). Minneapolis: University of Minnesota, Department of Psychology, Computerizing Adaptive Testing Laboratory.

Mazzeo, J., & Harvey, A. L. (1988). The equivalence of scores from automated and conventional educational and psychological test: A review of the literature (College Board report No. 88-8, TS RR. No. 88-21). NY: College Entrance Examination Board.

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss. (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing (pp. 223-236)*. New York: Academic Press.

McKinley, R.L., & Reckase, M.D.(1980). Computer applications to ability testing. *Association for Educational Data Systems Journal, 13*, 193-203.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114(3)*, 449-458.

Moreno, K.E., Wetzel, C.D., McBride, J.R., Weiss, D.J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. *Applied Psychological Measurement, 8*, 155-163.

Neal, V.A. (1991). *Comparing COMPUPASS with a paper and pencil version considering gender, computer experience, attitude toward computers, and test-taking anxiety*. Unpublished thesis. Texas Woman's University.

Olsen, J.B., Maynes, D.D., Slawson, D., & Ho, K.(1986). *Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement.* Paper presented at the Annual American Educational Research Association (6th, San Francisco, CA, April 16-20, 1986).

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31(3)*, 200-219.

Powers, D.E. & O'Neill, K. (1992). *Inexperience and anxious computer users: Coping with a computer-administered test of academic skills. The Praxis Series: Professional assessments for beginning teachers*. Educational Testing Service (ETS-RR-92-75), Princeton, N.J.

Ramsey, P.H. (1982). Empirical power of procedures for comparing two groups on *p* variables. *Journal of Educational Statistics, 7* (2), 139-156.

Raudenbush, S.W. (1994). Random effects models. In H. Cooper & L. Hedges (Eds.) *The handbook of research synthesis (pp. 301-321)*. New York: Russell Sage.

Raudenbush, S.W., Becker, B.J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111-120.

Rocklin, T., & O'Donnell, A.M. (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315-319.

Roose, L.L., Plake, B.S., & Wise, S.L. (1992). The effects of feedback in computerized adaptive and self-adapted test. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1993). Parametric measures of effect size. In H. Cooper & L. Hedges (Eds.) *The handbook of research synthesis (pp. 231-260)*. New York: Russell Sage.

Rosenthal, R. & Rosnow, R.L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. New York: Cambridge University Press.

Rosenthal, R. & Rubin, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99(3)*, 400-406.

Schaeffer, G.A., Reese, C.M., Steffen, M., Mckinley, R.L. & Mills, C.N. (1993). *Field test of a computer-based GRE general test*. GRE Board Report No. 88-08P. Educational Testing Service. Princeton, N.J.

Schnipke, D. L. (1995). Assessing speededness in computer-based tests using item response times. Paper presented at the Annual Meeting of the national Council on Measurement in education.

Siskind, T.G., Andrews, E.C., & Kovas, E.(1992). *The instructional validity of computer administered tests*. Paper presented at the Annual Meeting of the Eastern Educational Research association. Hilton Head, SC.

Smith, M.L., Glass, G.V., & Miller, T.I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Stone, G.E., & Lunz, M.E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education, 7(3)*, 211-222.

Strube, M.J. (1985). Combining and comparing significance levels from nonsignificance hypothesis tests. *Psychological Bulletin, 97(2)*, 334-341.

Traze, S. M., Elmore, P.B. & Pohlmann, J.T. (1992) Correlational meta-analysis: independent and nonindependent. *Educational and Psychological Measurement, 52*. 879-888.

Tyler, R.W., & White, S.H. (1979). Testing, teaching, and learning. U.S. Department of Health, Education and Welfare and National Institute of Education.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions, OTA-SET-519*. Washington, DC: U.S. Government Printing Office.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290-38315.

Vicino, F.L. & Hardwicke, S.B. (1984). *An evaluation of the utility of large scale computerized adaptive testing*. Paper presented at the meeting of American Educational Research Association, New Orleans.

Vispoel, W. P., & Coffman, D. D. (1994). Computerized-adaptive and self-adaptive music listening tests: Psychometric features and motivational benefits. *Applied Measurement in Education, 7(1)*, 25-51.

Vispoel, W.P., & Coffman, D. (1992). Computerized adaptive testing of music-related skills. *Bulletin of the Council for Research in Music Education, 41*, 111-136.

Vispoel, W.P., Wang, T., de la Torre, R., Bleiler, T., & Dings, J. (1992). *How review options and administration modes influence scores on computerized vocabulary*

tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for Testlets. *Journal of Educational Measurement, 24*, 185-201.

Ward, T. J., Hooper, S. R., & Hannafin, K. M. (1989). The effect of computerized tests on the performance and attitudes of college students. *Journal of Educational Computing Research, 5(3)*, 327-323.

Weiss, D.J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53(6)*, 774-789.

Weiss, D.J., & Kingsbury, G.G.(1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Wilder, G., Mackie, D., & Cooper, J. (1985). Gender and computers: Two surveys of computer-related attitudes. Sex Roles, 13, 215-228.

Wise, S.L., & Plake, B.S. (1989). Educational Measurement: Issues and Practice, 8(3), 5-10.

Wise, S.L. & Plake, B.S. (1989a). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, 5-10.

Wise, S. L. & Plake, B. S. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counseling Development, 23*, 3-10.

Appendix A

Primary studies

Baghi, H., Gabrys, R., & Ferrara, S. (1991). *Applications of computer-adaptive testing in Maryland.* Paper presented for the Annual Meeting of the American Educational Research Association, Chicago, April.

Baird, W.E. & Silvern, S.B. (1992). Computer learning and appropriate testing: A first step in validity assessment. *Journal of Research on Computing in Education, 25(1),* 18-27.

Bejar, I.I., & Weiss, D.J. (1978). *A construct validation of adaptive achievement testing.* Research report 78-4. Psychometric Methods Program, Department of Psychology, University of Minnesota.

Blackmore, L.M. (1986). *Computerized, computerized adaptive and pencil-and-paper test administration: A comparative study in a high school.* Unpublished dissertation. Pepperdine University.

Bugbee, A.C. Jr. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23(1),* 87-100.

Chin, C, H., J.S.D & Conry, R.F.(1991). Effects of computer-based tests on the achievement, anxiety, and attitudes of grade 10 science students. *Educational and Psychological Measurement, 51,* 735-745.

Dillon, G.F. (1992). *A comparison of traditional and computerized test modes and the effect of computerization on achievement test performance.* Unpublished dissertation. The Temple University.

Dimock, P.H., & Cormier, P. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement an Evaluation in Counseling and Development, 24,* 119-126.

Eaves, R.C., & Smith, E.(1986). The effect of media and amount of microcomputer experience on examination scores. *Journal of Experimental Education, 5,* No.1, Fall, 23-26.

Eignor, D.R. (1993). *Derining comparable scores for computer adaptive and conventional tests: An example using the SAT.* Paper resented at the Annual Meeting of the National Council on Measurement in Education. Atlanta, GA.

English, R. A, Reckase, M.D., & Patience, W.M.(1977). Application of tailored testing to achievement measurement. *Behavior Research Methods & Instrumentation, 9(2),* 158-161.

Federico, P. (1989). *Computer-based and paper-based measurement of recognition performance.* Navy Personnel Research and Development Center (NPRDC-TR-89-7).

Federico, P., and Liggett, N.L. (1989). *Computer-based and paper-based measurement of semantic knowledge.* Navy Personnel Research and Development Center, NPRDC-TR-89-4.

Glowacki, M.L., McFadden, A., & Price, B.J. (1995). *Developing computerized tests for classroom teachers: A pilot study.* Paper resented at the Annual Meeting of the Mid-South Educational Research Association. Biloxi, MS.

Harrel, T.H., Honaker, M., Hetu, M., and Oberwager, J. (1987). Computerized versus traditional administration of the Multidimensional Aptitude Battery-Verbal Scale.

An examination on reliability and validity. *Computers in Human Behavior, 3,* 129-137.

Henly, S.J., Klebe, K.J., & Mcbride, J.R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement, 13,* 363-371.

Hoffman, K.I., and Lundberg, G.D. (1976). A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement, 36,* 791-809.

Horton, S.V., & Lovitt, T.C. (1994). A comparison of two methods of administrating group reading inventories to diverse learners. *Remedial and Special Education, ?,* 378-390.

Huba, G.J. (1988). Comparability of traditional and computer Western Personnel Test (WPT) versions. *Educational and Psychological Measurement, 48,* 957-959.

Kim, J., & McLean, J.E. (1995). *The influence of examinee test-taking motivation in computerized adaptive testing.* Paper resented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Kovac, R.M.N. (1990). *The effects of computerized selection tests on job applicant performance.* Unpublished dissertation. DePaul University.

Kuan, T.H. (1991). *A comparison of paper-and-pencil, computer-administered, computerized feedback, and computerized adaptive testing.* Unpublished dissertation. Mississippi State University.

Lee, J.A. & Hopkins, L. (1985). *The effects of training on computerized aptitude test performance and anxiety.* Paper presented at the Annual Meeting of the Eastern Psychological Association (56th, Boston, MA).

Lee, J.A., Moreno, K.E., and Sympson, J.B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement, 46,* 467-474.

Legg, S. M., & Buhr, D.C.(1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, Summer,* 23-27.

Legg, S. M., & Buhr, D. (1990). *Investigating differences in mean score on adaptive and paper and pencil versions of the College Level Academic Skills Reading Test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Boston, MA. April.

Legg, S. M., & Buhr, D. (1987). *Final report: Feasibility study of a computerized test administration of the CLAST.* Institute foe Student Assessment & Evaluation, University of Florida.

Llabre, M.M., Clements, N.E., Fitzhugh, K.B., Lancelotta, G.L., Mazzagatti, R.D., & Quinones, N. (1987). The effect of computer-administered testing on test anxiety and performance. *Journal of Computing Research, 3(4),* 429-433.

Lunz, M.E. & Bergstrom, B.A. (1991). Comparability of decisions for computer adaptive and written examinations. *Journal of Allied Health,* 15-23.

Lunz, M.E. & Bergstrom, B.A. (1995). *Equating computerized adaptive certification examination: The Board of Registry Series of Studies.* Paper resented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

McDonal, J. (1988). *An analysis of children's performance on computer and paper and pencil administered assessments of whole number computation skills.* Unpublished dissertation. University of Washington.

Moon, O. (1992). *An application of computerized adaptive testing to the Test of English as a Foreign Language.* Unpublished dissertation. The State University of New York at Albany.

Muhlestein, A.L. (1991). *Comparison of the standard and computerized versions of the Collefe Level Examination Program general examination in English composition.* Unpublished dissertation. Utah State University.

Neal, V.A. (1991). *Comparing COMPUPASS with a paper and pencil version considering gender, computer experience, attitude toward computers, and test-taking anxiety.* Unpublished thesis. Texas Woman's University.

Olsen, J.B. (1990). Applying computerized adaptive testings in school. *Measurement and Evaluation in Counseling and Development, 23*, 31-38.

Parshall, C. & Kromrey, J.D. (1993). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Peterson, L., Johannsson, V., and Carlsson, S.G. (1996). Computerized testing in a hospital setting: Psychometric and psychological effects. *Computers in Human Behavior, 12(3)*, 339-350.

Power & Oneil (1992). *Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. The Praxis Series: Professional assessments for beginning teachers.* ETS-RR-92-75. Educational Testing Service, Princeton, N.J.

Rudy-Baese, P.A. (1979). *Computer or paper-pencil: A comparison of testing methods for non-standardized academic tests.* Unpublished dissertation. Marquette University.

Russell, M. & Haney, W. (1996). *Testing writing on computers: Results of a pilot study to compare student writing test performance via computer or via paper-and-pencil.* Paper presented at the Mid-Atlantic Alliance for Computers and Writing Conference.

Sorensen, H. B. (1985). Cognitive ability tests. *The Monitor, Nov./Dec.* 22-26.

Spray, J.A., Ackerman, T.A., Reckase, M.D., & Carlson, J.E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement, 26*, 261-271.

Tollefson, N. (1978). A comparison of computerized and paper-and-pencil formative evaluation. *College Student Journal*, 103-106.

Van de Vijver, F.J. & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852-859.

Vogel, L.A. (1994). Explaining performance on P&P versus computer mode of administration for the verbal selection of the graduate record exam. *Journal of Educational Computing Research, 11(4)*, 369-383.

Ward, B.C. (1994). *Student and teacher attitudes concerning computer adaptive testing methods in a middle school setting*. Unpublished Dissertation. The University of Maryland.

Ward, T.J., Hooper, S.R., and Hannafin, K.M. (1989). The effect of computerized tests on the performance and attitudes of college students. *Journal of educational Computing Research, 5(3)*, 327-333.

Watkins, M.W., & Kush, J.C.(1988). Assessment of academic skills of learning students with classroom microcomputer. *School Psychology Review, 17*, No.1, 81-88.

Wise, S.L., Barnes, L.B., Harvey, A.L., and Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education, 2(3)*, 235-241.